Computerized System to Aid Deaf Children in Speech Learning

Rodrigo Jardim Riella*, André Guilherme Linarth, Lourival Lippmann Jr., Percy Nohama
Programa de Pós-Graduação em Informática Aplicada/Pontifícia Universidade Católica do Paraná
Rua Imaculada Conceição 1155, 80215-901, Curitiba, PR, Brasil
percy@ppgia.pucpr.br

Abstract - **This paper describes a voice analyzer, whose purpose is deaf children's assistance in the process of speech learning. The processing of the user's speech signal is performed in real time in order to get an instantaneous feedback of the result of speech training. The aim of this analyzer is not to find the distinction between spoken words, main objective of a speech recognizer but to calculate a level of correctness in the toggle of a specific word. Voice signal analysis was developed through a digital signal processor (DSP), applying spectral analysis processes, extraction of voice's formants, adaptation of formants to the standard levels in domain of time and frequency and statistical matching of the acquired speech signal and the standard one, resulted from training. After calculating the correctness coefficient, the system goes off a visual feedback to the user in the form of a graphical animation, in accordance with the matching ratio, that will determine the progression on speech training.**

Keywords - **Speech impairment, digital signal processing, rehabilitation aid, deafness.**

I. INTRODUCTION

Data of the Brazilian Institute of Geography and Statistics (IBGE) indicate that the number of auditory impaired people reaches 1,5% in Brazil [1], which represents a number greater than 240 thousand persons.

People with auditory losses have difficulties and limitations with their oral communication that can cause from light to total loss of information.

Speech learning is strongly dependent of the auditory feedback because the speaker always hears the sound he makes by himself. Therefore, not only hearing performance but his talking is also damaged. When a person is learning how to articulate a new word, the perceived sound is compared to previous samples generated by others. So, the comparison induces a reduction of errors and an improvement of speech articulation.

The objective of presented system is the evaluation of deaf children articulatory process, looking for a funny way of teaching and training the correct articulation of words. The method dynamics is provided by the creation of a visual return, such as in a video game, where the evolution depends on the correct pronunciation of proposed words.

II. METHODOLOGY

Speech evaluation is realized extracting the three main formants (resonance frequencies of buccal, nasal and larynx cavities) that can be used as a parameter either to recognition or to evaluate words articulation. Formants are acquired from voice signal through three digital filter banks, centered on the predominant spectral bands.

Three vectors are extracted to represent the spectral variation of the position of each formant along the time, using a time resolution of 5 ms. Those vectors can be used either to generate pattern data for comparison or evaluation, which are obtained calculating the linear correlation coefficients among the signal under test and the patterns. The result informs how near the evaluated vector is from the standard one.

The magnitude of this coefficient defines a reference level to the computer in order to trigger a graphical animation presented to the user as a means of visual feedback. The evolution of the animation scene is a measure of the advance in articulatory training.

The architecture of the articulation patterns evaluator can be divided in three modules (formants extraction, temporal normalization and mathematical ones), which work sequentially and defining the phases of the process of the parameter's extraction, normalization and comparison.

2.1. Formants extractor module

Three filter banks were created [3] considering the spectral bandwidth where children's voice formants are found, as illustrated in the Table I. The three banks are composed of 29 FIR filters of 100 taps, using Hamming window [4]. This method was chosen as a consequence of temporal resolution of 5 ms in the formants position variation, because using an spectral analysis via FFT, sampling at 10.8 kHz, it would result on only 64 resolution points of the whole analyzed spectrum, caused by the time x frequency resolution paradox.

Table I. Children's Voice Formants Localization [5].

Vogal	F_1		F_2		F_3	
	μ	σ	μ	σ	μ	σ
/a/	1086	82,84	1721	195,91	2873	246,71
/e/	902	83,39	2606	187,26	3243	187,76
/ɛ/	698	76,30	2825	288,60	3637	244,19
/i/	465	81,70	3176	216,36	3980	212,27
/ɔ/	913	100,97	1371	81,02	2793	216,74
/o/	682	73,35	1295	103,55	2823	218,07
/u/	505	93,32	1350	130,24	2667	289,40

μ = mean position , in Hz ; σ = standard deviation

In order to minimize the interference of the environment noise over the speech, taken in the beginning and in end of articulated word, it was implemented an automatic microphone turn on - turn off algorithm, using a level

*Undergraduate student of the Electrical Engineering Course at CEFET-PR, Curitiba, Paraná, Brazil.

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25OCT2001	N/A	-

Title and Subtitle Computerized System to Aid Deaf Children in Speech Learning	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Programa de Pós-Graduação em Informática Aplicada/Pontifícia Universidade Católica do Paraná Rua Imaculada Conceição 1155, 80215-901, Curitiba, PR, Brasil	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 25-28 OCT 201, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

detector for the entrance signal, compared with the environment noise, which triggers the starting process if the input signal is higher than the threshold and ends the process if this signal becomes lower than that threshold during a certain period of time. By this way, the formants extractor module turns on automatically when the input sound reaches the reference level, passes it though the three filter banks and stores the output sum of each filter during the 5 ms resolution time.

After this interval, the system tests all filter's outputs, looking for the highest value of each bank and storing the filter values in three vectors which represent the position variation of each formant along the time. The formants extractor module block diagram is represented in the Fig. 1.

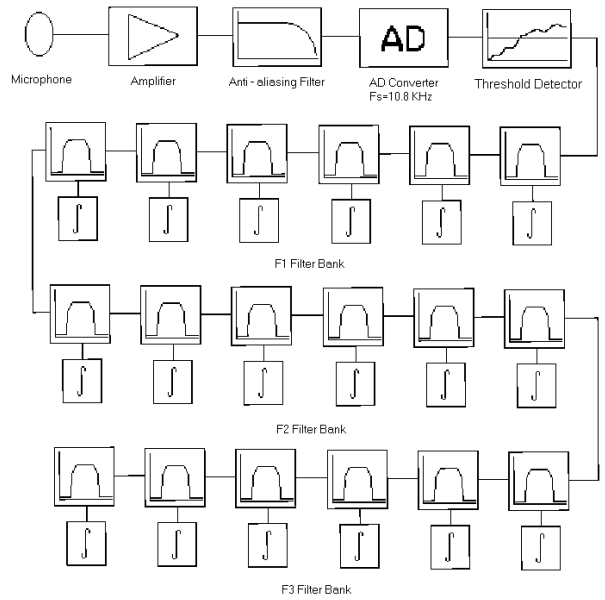

Fig. 1. – Simplified block diagram of formants extractor module. The integration period is 5 ms, which can be adjusted.

The system stays in loop during the period of word's articulation, obtaining all the vectors of the complete word. After detecting the end of a word articulation, the system manages the temporal normalization and mathematical modules (Fig.2).

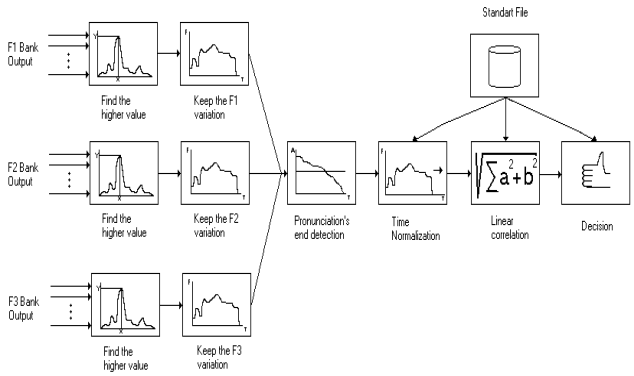

Fig. 2 – Block diagram of time normalization, mathematical and decision modules.

2.2 Temporal Normalization and Mathematical Modules

In consequence of applying a correlation-based algorithm to define the correctness index, the obtained vectors need to be normalized in order to display a standard length, avoiding to penalize the words correctly articulated but during different periods of time.

The method developed to get that normalization is based on the insertion of samples into vectors until they have similar length, without changing so much its variation, improving the linear correlation coefficient among the vectors if they are similar, and maintaining this coefficient very low if they are different.

To calculate the proximity index among the obtained vectors and their standard ones, it is determined the linear correlation coefficient among them. This coefficient is mathematically solved by equation (1) [2].

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \quad (1)$$

The magnitude of this coefficient can vary between –1 to +1. Magnitude +1 indicates that the vectors are completely correlated in a positive way and the counterpart, magnitude –1, means that the vectors are completely correlated in a negative way. If the linear correlation coefficient is zero, it indicates the vectors are completely uncorrelated.

Since this system was designed to analyze the variation of the spectral position of the formants as a function of time, negative values of "r" are not interesting, expressing an error of meaning. For this application, the most interesting values are the ones nearest +1.

The resultant magnitude of the linear correlation coefficient is used either (1) to create a standard file, which contains the statistics of this coefficient's variation during the training of new words, or (2) to verifies through the decision algorithm if the obtained distribution are inside the range obtained during the training session and now contained inside the standard file.

2.3. User's Interface

In order to perform the visual feedback to the user of the system (called VOXsis), it was developed a program in C++ Builder 5.0 language, looking for creating gradual challenges through animations and games that makes VOXsis pleasant and motivating for children daily practice and training.

These challenges can be programmed by the phonologist who has the control of what kind and which words will be articulated by each user during his training session, and what sequence it must be applied. VOXsis front page can be viewed in Fig.3.

A new input register based on child's name automatically adjusts the game's features setup, for every user. To do this, the phonologist can program the desired options for everyone through the configuration menu. These options are protected by passwords in order to prevent modification by not authorized persons. By this way, the

child can operate the program by herself once the system has been previously configured by the phonologist for child's training. Therefore, it only needs to enter its register name and to click on the corresponding icon to begin a new session, that initiates the routine of programmed challenge for daily speech training. Beyond the graphical animation, the user visualizes a bar that indicates the level of proximity between its word articulation and the stored training pattern, being able to see the resultant graphics of its word articulation, and making possible to observe the actual matching related to the patterns, simply selecting graphics icon.

Fig. 3 – VOXsis front page, where the user can choose between start game or go to user setups (protected by passwords).

There is also an option to visualize pictures and animations of phoneme's pronunciation, for training of labial reading. The phonologist can easily add pictures to the system, just coping the files to the specific directory, because VOXsis accept pictures and draws in the bitmap format.

III. RESULTS

In order to perform preliminary tests for the speech analysis system, it was yielded audio signals created using an audio edition software. Their amplitude, duration and environment noise levels were changed for measuring the system's performance.

The first generated test signals were frequency sweepings by which the performance of the filter banks was investigated. Since banks 2 and 3 have a crossing on their frequency bands due the localization of formants 2 and 3 depending on the word sharpness, it was created only one sweep for investigating those two banks sequentially. After that, the signal was mixed to the sweep for verifying bank 1. By this way, the desired result was a sequential action on the output of the filter banks, represented by a "staircase", which really was confirmed. It can be seen on Fig. 4, where filter bank 1 was put into action during the signal application, while banks 2 and 3 were put into action sequentially between them but in a parallel mode with respect to bank 1.

Since the yielded original test signal had 300 ms, the system was trained with signals having time variation between 100 to 500 ms. It was verified that for time variations up to 50% around the original signal, the system presented little variation in the linear correlation coefficient, with minimum values around 0.85, which doesn't degenerate the recognition ratio. Increasing the variation, this coefficient's value starts to fall drastically, mainly due to distortion added by the time normalization algorithm, which loses its characteristics with the increasing of its conformation, and degenerating the original signal. It is extremely recommended that the articulated words duration does not exceed 50% of the training's duration in order to avoid this problem. That test was repeated with the same sweepings of 1 s duration and it was noticed the same result, since the limitation of the word's length is about 2 s.

Varying the amplitude of the input signal, it was not found problems, since the analysis was developed in the frequency domain. The system had problems only in the recognition ratio when the input signal amplitude passed over the quantization A/D converter levels.

When adding white noise to the signal, it was not noticed problems in the recognition mode while this noise was held in levels lower than 3 dB over the original signal. To higher levels, the created distortions change the signal causing problems in the recognition block.

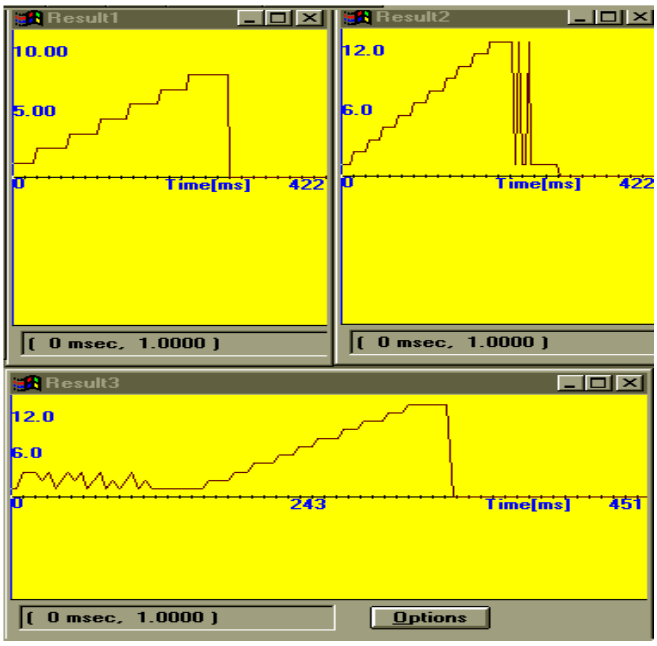

Fig. 4. – Formants extractor module's results to the test sweep, with duration of 300 ms. Each graphic represents a filter bank, and each axis represents the time and the number of the excited filter.

The recognizing tests were made in three steps: first were changed the ranges of frequency of the sweepings and delivered to the system. After that, tests with multifrequential signals were made, creating sequences that were applied to the recognition module. Finally, the tests were made with speech signals, in order to analyze the performance of the prototype.

The system presented optimal results with the recognizing tests using sweepings, where the changed signals were not recognized in 100% trials, presenting on the other hand 100% of recognition ratio for the trained sweepings. For multifrequential signals, the system presented 95% of correctness in the recognition ratio for trained data in spite of varying the amplitude, duration and the environment noise level.

The speech signal tests are still being executed, however, with the preliminary results, using men voices, it reached around 65% in the recognition ratio. For those tests,

it was selected brazilian portuguese words such as “alô”, “teste” and “um”.

IV. DISCUSSION

The tests of speech signals still have to be completed, because the kind of test applied did not evaluate the system’s behavior completely, because the filter banks are centered for the spectral regions of children’s formants localization, which are different of the adult formants localization. Those tests were performed only to make a preliminary prototype evaluation. The prototype still presents some limitations such as word’s duration, which can’t be over 50% of variation related to the reference period, requiring a phonologist’s helping when starting VOXsis, who has to determine the time duration of each word in use for every specific user, and that is available in the configuration menu. Some adjusting conditions and setting levels mainly related to different microphones that can be implemented. These equalizing algorithms are still in development and has been incorporated to the prototype as soon as they are functioning satisfactorily.

VOXsis software package still have only two graphic animations to user’s visual feedback. This number will be increased to improve the children stimulation in the use of the system, but it is satisfactory to evaluate the prototype.

V. CONCLUSIONS

Although the applied tests have not presented a complete view of the system’s functioning, they demonstrate that VOXsis will present a satisfactory performance.

Only after solving the necessary fittings, it will be performed the final tests with deaf children. Then, VOXsis have to be evaluated by phonologists who will provide technical data for the conclusion of the project and the way it can be inserted into speech learning programmes.

A complete evaluation and proper development of VOXsis final version will depend on making prototypes available to schools, clinics and home care.

ACKNOWLEDGMENTS

The authors are grateful to CNPq, Fundação Araucária and LACTEC for their support during the development of this research project.

REFERENCES

[1] IBGE, "Instituto Brasileiro de Geografia e Estatística. Tipos de Deficiências". Available in the Internet: http://www.sidra.ibge.gov.br/bda/tabela/listabl.asp?e=l&c=138

[2] W.H. PRESS, S. A. TEUKOLSKY, W. T. VETTERLING and B. P. FLANNERY, "Numerical Recipes in C: The Art of Scientific Computing", 1988, Cambridge University Press, USA.

[3] L. RABINER and B. JUANG, "Fundamentals of Speech Recognition", Prentice Hall, 1993, UK.

[4] A.G. REICHEL, "Estudo e Desenvolvimento de um sistema de Medição e Análise de Ruído Acústico em Ambientes Industriais para Auxílio na Identificação de Perdas Auditivas", Dissertação de Mestrado, CPGEI / CEFET-PR, 1999, Curitiba, Brazil.

[5] I. RUSSO and M. BEHLAU, "Percepção de Fala: Análise Acústica do Português Brasileiro", Ed. Lovise, 1993, São Paulo, Brazil.